DOCUMENT RESUME

ED 283 836 TM 870 338

AUTHOR	Gaines, Margie L.
TITLE	Evaluating Magnet Schools Effectively: Challenges and Cautions.
INSTITUTION	Austin Independent School District, Tex. Office of Research and Evaluation.
SPONS AGENCY	Department of Education, Washington, DC.
REPORT NO	AISD-86.37
PUB DATE	Apr 87
GRANT	165AH50033
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
AVAILABLE FROM	Margie L. Gaines, Office of Research and Evaluation, 6100 Guadalupe, Austin, TX 78752.
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Academic Achievement; *Achievement Gains; Curriculum; *Evaluation Methods; Factor Analysis; Factor Structure; High Schools; Item Analysis; Latent Trait Theory; *Magnet Schools; Science Curriculum; Science Tests; *Secondary School Science; *Standardized Tests; Test Validity
IDENTIFIERS	Austin Independent School District TX; Iowa Tests of Basic Skills; *Science Academy of Austin TX; *Tests of Achievement and Proficiency

ABSTRACT

Magnet schools have become increasingly popular as alternatives to traditional curriculum offerings and as tools to promote voluntary desegregation in urban school districts. Most evaluations of magnet schools have been concerned with monitoring objectives and compliance with regulations. Few reports assess the impact of magnet curricula on student achievement. The evaluation of achievement in magnet programs cannot be interpreted when there are no comparisons of magnet students with nonmagnet students or with scores district-wide. The Austin (Texas) Independent School District implemented the Science Academy of Austin, a high school science, mathematics, and computer technology magnet program in 1985. At the end of the first year the tenth-grade magnet students' achievement gains did not significantly out-gain their district-wide counterparts in science. Two reasons were hypothesized: (1) the Tests of Achievement and Proficiency (TAP) Science test did not address the curriculum taught to tenth-grade magnet students; and (2) the TAP Science test was not measuring what the test publishers said the test should measure. A factor analysis of the TAP Science test gave support to both hypotheses. Evaluators should examine their implicit assumptions about the appropriateness of using results of standardized tests to evaluate special programs. (BAE)

ED283836

EVALUATING MAGNET SCHOOLS EFFECTIVELY: CHALLENGES AND CAUTIONS

by

Margie L. Gaines

Office of Research and Evaluation
Austin Independent School District
Austin, Texas

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

M. L. Gaines

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DISCLAIMER

The Magnet Schools Assistance Program in the Austin Independent School District was funded by a grant from the Department of Education, No. 165AH50033, Federal Identification No. 74-6000006-3. The opinions expressed are those of the author and do not necessarily reflect the position or policy of the Department of Education, the Austin Independent School District, or the Office of Research and Evaluation. No official endorsement by any of these organizations should be inferred.

Paper presented at the American Educational Research Association Annual Meeting, Washington, D.C., April 20-24, 1987.

Requests for copies and correspondence should be directed to the author at: Office of Research and Evaluation, 6100 Guadalupe, Austin, TX 78752

Publication Number 86.37

TM 870 338

EVALUATING MAGNET SCHOOLS EFFECTIVELY: CHALLENGES AND CAUTIONS

INTRODUCTION

Magnet schools are becoming increasingly popular as educational alternatives to traditional curriculum offerings and as tools for promoting voluntary desegregation in urban school districts. Often, the principal objectives of magnet programs are oriented toward meeting ethnic distribution goals or increasing enrollment in underenrolled schools. Hence, the evaluations of these programs are usually objective-driven and focus on student characteristics, minority recruitment and retention, parent and community involvement, and whether the programs offer unique educational and interracial experiences (e.g., Stanley, 1984). The evaluation of federally funded programs requires attention to compliance with federal guidelines, student profiles, how well the programs contributed to reducing minority group isolation, and to whether local objectives specified in the grant proposal were met.

While program staff and evaluators have been concerned with monitoring objectives and compliance with regulations, little attention has been focused on how well magnet programs promote student achievement through the enhanced curricula. Only a few reports have attempted to assess the impact of magnet curricula on student achievement (e.g., Abadzi & Dunkins, 1984; Zepeda, 1986).

Unfortunately, these reports attempt to compare the achievement of magnet students to district and national norms without taking into

consideration the students' previous level of achievement. Some programs have selective admission criteria and admit only high-achieving or above-average students (such as the program in Fort Worth, TX). Appropriate evaluations of student achievement in these cases is further complicated because the range of ability is restricted to the upper percentiles. In such situations, it is more likely that the students in the magnet program are maintaining their lead rather than making larger gains than other students in the district.

The evaluation of achievement in magnet programs is uninterpretable when there are no comparisons of magnet students with nonmagnet students or with scores districtwide. While the evaluation of magnet programs in Dallas (Zepeda, 1986) provides pre- and posttest scores, district and national comparative data are lacking. Until evaluators report magnet program achievement data in the context of appropriate comparison groups, it is difficult to assess accurately the impact of a magnet program on student achievement.

AUSTIN'S EXPERIENCE.

During the 1985-86 school year, the Austin Independent School District in Austin, Texas implemented the Science Academy of Austin, a high school science, math, and computer technology magnet program. Students were selected for admission based upon their standardized test scores, which had to be above the 50th percentile on all subtests. The Iowa Tests of Basic Skills (ITBS) is used for eight-grade applicants and the Tests of Achievement and Proficiency (TAP) is used for ninth-grade

applicants. The median percentile scores of students accepted into the Science Academy were more than 30 points above students districtwide in math, science, and reading.

Figure 1

MEDIAN PERCENTILE SCORES OF EIGHTH- AND NINTH-GRADE STUDENTS ACCEPTED INTO THE SCIENCE ACADEMY IN FALL, 1986

	ITBS (8TH GRADE)		TAP (9TH GRADE)	
	Academy	District	Academy	District
Reading:	91	54	90	54
Mathematics:	90	54	88	55
Science:	*	*	90	54

* Austin ISD does not administer the ITBS science subtest.

Students at the upper percentile ranges typically make larger gains in grade equivalent scores during an academic year than do students in the 50th percentile range. To compare pre- to posttest gains of magnet students to the districtwide median gains without statistically controlling for pretest scores would be misleading. That the Science Academy students were high-achievers necessitates taking into account their level of achievement upon entering if comparisons with similar high-achieving students districtwide were to be meaningful.

ASSESSING CURRICULUM IMPACT: TESTING PROBLEMS ENCOUNTERED.

The Science Academy faced the challenge of assessing the influence of the math and science curriculum on the magnet students after just one year in the program. Both ninth- and tenth-grade magnet students took at

least two periods of science each day, thereby completing a year-long course in one semester. A few students also took a third period of science. Ninth-grade Science Academy students took two periods of biology, and some also took chemistry or physics. The tenth-grade science magnet curriculum included chemistry and physics, although some students also took biology.

By the end of the year, ninth-grade magnet students had a median mathematics gain of 5.30 years (in grade equivalent scores). Ninth-grade students districtwide had a median gain of 1.63 years in mathematics. Tenth-grade magnet students had a median gain of 2.00 years in mathematics but less than one year (0.7 grade equivalent) in science. On the other hand, tenth-grade students districtwide had a median gain of 2.35 years in math and over two years in science. Grade equivalent scores alone do not reveal whether the gains made by the magnet students were as large as might be expected given that they already were scoring in the top percentiles and were receiving enriched, accelerated instruction.

The Office of Research and Evaluation annually assesses achievement gains through a linear regression model that controls for students' background characteristics, including sex, ethnicity, and low-income status, as well as their level of achievement from the previous year. The model predicts a score for each student and then compares the student's actual score with the expected level of achievement. In this way, the achievement level of the magnet students as a group could be compared to their high-achieving counterparts who were not in the program

to get an idea of the impact of the curriculum, which was the main difference once the influence of all other variables had been controlled.

The results of the regression analyses showed that while the Science Academy students exceeded their predicted levels of achievement in all areas, not all the differences were significantly larger than the gains made by similar students who were not in the program. The tenth-grade magnet students did not significantly out-gain their districtwide counterparts in science, although their gains in math were significantly larger than students districtwide.

Reasons were hypothesized for the lack of significant science gains for tenth-grade students despite the accelerated, enriched instruction at the Science Academy. A primary hypothesis was that the TAP Science test (Level 16), which is given to all tenth-grade students districtwide, did not address the curriculum taught to tenth-grade magnet students. Hence, it was thought that a curriculum-test mismatch was one plausible explanation. A second hypothesis was that the TAP Science test was not measuring what the test publishers said the test was designed to measure, that is, a variety of areas of science content knowledge and skills.

In order to investigate the hypotheses, the TAP test manual was scrutinized for information on the construction of the science test and on the content and skills that it intends to measure. Rather than assume the appropriateness of the standardized test, a statistical approach was undertaken for determining whether the test was adequate as a tool for evaluating the impact of an enriched curriculum on above-average students.

The focus here is on the use of the test as an evaluation tool with a special group and not on the psychometric adequacy of the test for measuring achievement.

FACTOR ANALYSIS AND ITEM ANALYSIS.

The TAP Science questions were factor analyzed to reveal the underlying structure of the test. Traditional item analyses also were performed to determine the difficulty of each item and the range of item difficulty across the test.

The Riverside Publishing Company reports that the TAP Science test classifies items by subject matter: Biology, Nature of Science, Earth and Space Science, Physics, and Chemistry. Items are also classified by the type of skill used in responding to items. The manual says the skills, the functional techniques for discovery and understanding, are: Knowledge, Application, Explanation, and Experimental Methods and Techniques.

According to the item classification scheme in the test manual, the 62 items of the tenth-grade science test break down into the following percentages of items covering each content and skill area. (Because of rounding, the percentages sum to 99.)

Figure 2

LEVEL 16 TAP SCIENCE SUBTEST ITEM CLASSIFICATION

CONTENT AREAS		SKILL AREAS	
Biology:	37%	Knowledge:	47%
Nature of Science:	29%	Application:	14%
Earth and Space:	27%	Explanation:	19%
Physics:	3%	Experimental Methods	
Chemistry:	3%	and Techniques:	19%

The responses from 3,627 students to the 62 items on the test were factor analyzed to investigate whether the underlying structure of the test supported the item classification scheme. A principal component analysis was performed on the item intercorrelation matrix using communalities equal to one in the principal diagonal and a scree test to determine the number of factors to extract. Three factors were retained and a varimax (orthogonal) rotation was done to find the factor loading of each item on the three factors.

RESULTS

The factor analysis resulted in three factors being extracted, with one main factor. The first factor had an eigenvalue of 9.86, the second had an eigenvalue of 2.05, and 1.78 for the third factor. Essentially, the test is unidimensional. The factor loadings of the 62 items of the tenth-grade TAP Science test (Level 16, sequential item numbers 19-80) are presented in Table 1.

The traditional item analysis revealed that the item difficulty, indicated by the proportion of students passing each item, ranged from a low passing rate of 34% to a high of 86% on an item. The median was a 55% passing rate. The majority of the items (69%) had a passing rate greater than 50%. Classical mental test theory considers a 50-50 split desirable. The item P-values (percent of students passing the item) and the item-total (uncorrected) correlations are presented in Table 2.

The first factor, which had 26 items loading on it, had a mean passing level of 66%. The second factor, with 19 items, had a mean passing rate of 52%, and factor three, with 16 items, had a mean of 47%.

On the average, the easiest items loaded on the first factor. Factor one contained items from the content areas of Biology, Nature of Science, and Earth and Space Science, and only one Chemistry item. All skill areas were represented but with a predominance of knowledge level items. Hence, it appeared that the first (and primary) factor might be labeled "General Science Knowledge."

DISCUSSION

The factor analysis results indicated fewer underlying factors than either the number of content areas or skill areas indicated by the test publishers. The traditional item analysis was performed to further elucidate the meaning of the underlying factors. Given that the first factor contained the items with the highest passing rates, the factor also could be an "easiness" factor.

The factor structure of the TAP science subtest supports the conclusion that the test measures general science knowledge. However, it is not as factorially complex as the item classification would suggest nor as complex as the test constructors intended. Thus, there was support for the hypothesis that the TAP Science subtest does not measure the factors it purports to measure. The item classification scheme appears to have face validity but not factorial validity. While this finding may not challenge the validity of the test as a measure of general science achievement, it does pose problems when using the test as an evaluation tool for a science magnet program.

Also, there was support for the hypothesis that there was a mismatch between the Science Academy tenth-grade science curriculum and the content of the test. There were very few chemistry or physics items--too few to reflect the advanced curriculum of the Science Academy, which emphasized chemistry at the tenth-grade level. Furthermore, there was an insufficient number of difficult items to adequately discriminate among students at the upper levels. It appears the test was designed to measure minimum to average levels of science knowledge.

As an evaluation tool for use with advanced science students, the TAP Science test is insufficient. These results highlight the need to scrutinize evaluation tools, especially when a standardized test is being used with special groups for evaluating program effects. Effective program evaluation depends upon the use of appropriate instruments. When the focus of an evaluation is on the effect of a specialized curriculum compared to the effects of the regular curriculum, and not necessarily on student achievement per se, then the appropriateness of the instrument and its match to the curriculum or program objectives is critical.

When a commonly used test fails to match the curriculum areas of instruction, alternatives must be sought. Because the TAP Science subtest at the tenth-grade level was not sufficiently sensitive to the effects of a nonstandard curriculum on nonstandard students, it does not follow that the test has no use in this situation. An alternate approach should be sought for evaluating the achievement of Science Academy students (in addition to, not instead of regular achievement testing).

One possibility that will be examined is using Item Response Theory (IRT) techniques to calibrate the items available on all levels of the test. Once the difficulty, discrimination, and guessing parameters of each item are known, a test may be designed from the existing items for use with the Science Academy students. A selected set of items could be used for obtaining estimates of each student's ability level which depends neither on the items nor on the original norming sample of students. Changes in ability level for each student (or for the group of magnet students) as a result of exposure to the magnet program curriculum could be measured each year. A control group of students for comparison could easily be found with similar demographic and achievement characteristics.

In conclusion, this study should stimulate evaluators to examine their implicit assumptions about the appropriateness of using results of standardized tests to evaluate special programs. Testing results are readily available to evaluators, yet they may not be reliable as an index of the program's effects on students. Furthermore, when comparing the achievement of program participants to regular students, the previous level of achievement must be used as a covariate, along with other demographic variables, so that comparisons between program participants and nonparticipants are statistically reasonable as well as interpretable. At times it is incumbent upon the evaluator to abandon assumptions and to investigate through statistical techniques the appropriateness and adequacy of their evaluation instruments.

REFERENCES

Abadzi H. and Dunkins D. (1984) A model for a magnet program which promotes both high achievement and voluntary integration. Paper presented at the American Educational Research Association Annual Conference, April, 1984. (ERIC Document Reproduction Service No. ED 244 041)

Riverside Publishing Company. (1982) Manual for School Administrators: Tests of Achievement and Proficiency. Chicago: Author.

Stanley, C., et al. (1984) Magnet schools: Ninth annual final report, 1983-84. (ERIC Document Reproduction Service No. ED 259 058)

Zepeda, R. A. (1986) Evaluation of 1985-86 Vanguard, Academy, and Magnet High School Programs (REA86-032-5). Dallas, TX: Dallas Independent School District, Department of Research, Evaluation and Audit.

Table 1

ITEM FACTOR LOADINGS

PRINCIPAL COMPONENTS ANALYSIS WITH VARIMAX ROTATION
ROTATED FACTOR PATTERN

ITEM NUMBER	FACTOR 1	FACTOR 2	FACTOR 3
33	0.56909	-0.13266	-0.02608
29	0.50338	-0.14179	0.07075
32	0.47817	-0.17055	0.07873
79	0.46718	-0.06596	0.15257
36	0.44885	-0.29586	0.25091
53	0.44362	-0.12567	0.15777
76	0.44281	-0.22546	0.33160
71	0.43732	-0.08725	0.25398
26	0.43064	-0.20706	0.04061
69	0.42867	-0.23780	0.34118
21	0.42161	-0.16169	0.01930
42	0.40103	-0.18325	0.11379
75	0.40044	-0.08371	0.19229
44	0.38487	-0.07539	0.15928
35	0.38292	-0.24082	0.07522
45	0.37228	-0.10409	0.12221
34	0.36958	-0.21658	0.17687
68	0.36136	-0.05492	0.10838
31	0.34183	-0.10629	0.12673
48	0.34167	-0.17490	0.15593
19	0.34113	-0.18285	0.14837
65	0.29981	-0.07550	0.23791
52	0.28813	-0.07119	0.11663
43	0.28578	-0.19322	0.26614
63	0.26732	-0.07858	0.24713
50	0.22258	-0.09120	0.15695
39	-0.20058	0.55712	-0.12623
54	0.03436	0.52583	-0.26640
41	-0.23182	0.50937	-0.09424
30	-0.07529	0.47239	-0.12556
64	-0.17192	0.47123	-0.07719
72	-0.17931	0.46607	-0.08723
25	-0.02977	0.46231	-0.17284
70	-0.10074	0.43865	-0.12549
23	-0.09439	0.43640	-0.15859
28	-0.25292	0.42552	-0.07028
20	-0.09395	0.42253	0.07422
61	-0.23338	0.41981	-0.01195
47	-0.06031	0.40131	-0.11355

Table 1 (cont.)

ROTATED FACTOR PATTERN

ITEM NUMBER	FACTOR 1	FACTOR 2	FACTOR 3
40	-0.30491	0.38651	0.11269
51	-0.24891	0.37700	-0.02996
66	-0.18346	0.36682	-0.03898
59	-0.12688	0.33998	-0.12747
78	-0.17558	0.33940	-0.00115
27	0.16968	0.17083	0.04862
46	0.31996	-0.33443	0.33385
55	-0.05636	-0.15211	0.64610
56	-0.04142	-0.17784	0.63639
60	0.21951	-0.10242	0.42623
74	0.02058	-0.03373	0.41265
58	0.13914	-0.01769	0.40577
67	0.20185	-0.24172	0.39907
37	0.24409	-0.19457	0.37077
62	0.24694	-0.19394	0.36475
38	0.21639	-0.32376	0.34939
80	0.30730	-0.09065	0.34208
73	0.32644	-0.14196	0.33127
49	0.11870	-0.08010	0.30384
77	0.12067	0.13930	0.30384
57	0.11359	0.00699	0.28155
24	0.10829	-0.02452	0.26193
22	0.11514	-0.04420	0.25981

VARIANCE EXPLAINED BY EACH FACTOR

FACTOR 1	FACTOR 2	FACTOR 3
5.332503	4.623753	3.747682

Table 2

TAP LEVEL 16 SCIENCE SUBTEST

ITEM PASSING RATE (P VALUE) AND ITEM-TOTAL CORRELATIONS

ITEM NUMBER	P VALUE	R(TOTAL)	SIGMA
19	.67	.4348	.469
20	.80	.3834	.401
21	.86	.4498	.347
22	.50	.2648	.500
23	.51	.4185	.500
24	.49	.2590	.500
25	.44	.4018	.496
26	.81	.4738	.396
27	.41	.0972	.491
28	.55	.4843	.497
29	.75	.4792	.431
30	.55	.4286	.497
31	.74	.3962	.438
32	.77	.4886	.423
33	.85	.4965	.360
34	.58	.4673	.494
35	.75	.4686	.431
36	.66	.5939	.474
37	.38	.4470	.486
38	.60	.5024	.489
39	.60	.5396	.490
40	.75	.4514	.436
41	.59	.5208	.492
42	.82	.4784	.384
43	.72	.4623	.450
44	.66	.4080	.473
45	.68	.3997	.466
46	.57	.5629	.495
47	.39	.3650	.488
48	.67	.4354	.471
49	.48	.3144	.500
50	.62	.3367	.486
51	.52	.4391	.500
52	.46	.3236	.499
53	.64	.4672	.479
54	.40	.4363	.489
55	.37	.3823	.482
56	.39	.4008	.488
57	.42	.2550	.494
58	.36	.3261	.479
59	.34	.3737	.475

Table 2 (cont.)

ITEM PASSING RATE (P VALUE) AND ITEM-TOTAL CORRELATIONS

ITEM NUMBER	P VALUE	R(TOTAL)	SIGMA
60	.43	.4318	.495
61	.60	.4656	.490
62	.52	.4787	.500
63	.51	.3815	.500
64	.53	.4658	.499
65	.55	.4056	.498
66	.48	.4052	.499
67	.61	.5063	.489
68	.62	.3851	.486
69	.57	.5981	.495
70	.44	.4224	.496
71	.55	.4958	.498
72	.53	.4745	.499
73	.57	.5013	.495
74	.44	.3024	.496
75	.64	.4602	.481
76	.55	.6018	.497
77	.39	.2332	.488
78	.56	.3900	.497
79	.73	.4841	.442
80	.44	.4570	.496